Regression models

INTD8065 Data Analysis for Cancer Research
María-Eglée Pérez and Luis Raúl Pericchi

Contents of this class

1. Correlation

2. Linear Models

3. Simple Regression Model
   - Estimation in the Simple Regression Model
   - Tests of Hypothesis on the Parameters of the Model
   - Analysis of Variance
   - Multiple Determination Coefficient (R^2)
   - Regression Diagnostics

4. Multiple Linear Regression
   - Hypothesis Testing for the Model Parameters
   - Multiple Determination Coefficient R^2
   - Prediction
   - Comparing non-nested models

Correlation

Consider a situation where we are interested in the statistical relationship between two random variables, designated Y_1, Y_2, in a population. Both variables are continuous and each sampling or experimental unit (i) in the population has a value for each variable, y_{i1} and y_{i2}.

One measure of the strength of a *linear* relationship between two continuous random variables is the *covariance*.

$$\sigma_{Y_1 Y_2} = E\left[(Y_1 - E(Y_1))(Y_2 - E(Y_2))\right]$$

Estimator based on sample data:

$$s_{Y_1 Y_2} = \frac{1}{n-1}\sum_{i=1}^{n}(y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)$$

One measure of the strength of a *linear* relationship between two continuous random variables is the *covariance*.

$$\sigma_{Y_1 Y_2} = E\left[(Y_1 - E(Y_1))(Y_2 - E(Y_2))\right]$$

Estimator based on sample data:

$$s_{Y_1 Y_2} = \frac{1}{n-1}\sum_{i=1}^{n}(y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)$$

One limitation of covariance is its dependency on the units of the two variables. A way of overcoming this limitation is standardizing the covariance by dividing by the standard deviations of the two variables, so that our measure of the strength of the linear relationship lies between -1 and 1. This is called (*Pearson* or *product-moment*) *correlation*

Population definition:

$$\rho_{Y_1 Y_2} = \frac{E\left[(Y_1 - E(Y_1))(Y_2 - E(Y_2))\right]}{\sigma_{Y_1}\sigma_{Y_2}}$$

Sample estimator:

$$r_{Y_1 Y_2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)}{s_{Y_1}s_{Y_2}}$$

One limitation of covariance is its dependency on the units of the two variables. A way of overcoming this limitation is standardizing the covariance by dividing by the standard deviations of the two variables, so that our measure of the strength of the linear relationship lies between -1 and 1. This is called (*Pearson* or *product-moment*) *correlation*

Population definition:

$$\rho_{Y_1Y_2} = \frac{E\left[(Y_1 - E(Y_1))(Y_2 - E(Y_2))\right]}{\sigma_{Y_1}\sigma_{Y_2}}$$

Sample estimator:

$$r_{Y_1Y_2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)}{s_{Y_1}s_{Y_2}}$$

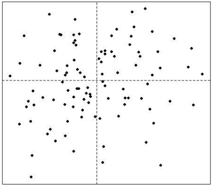
One limitation of covariance is its dependency on the units of the two variables. A way of overcoming this limitation is standardizing the covariance by dividing by the standard deviations of the two variables, so that our measure of the strength of the linear relationship lies between -1 and 1. This is called (*Pearson* or *product-moment*) *correlation*

Population definition:

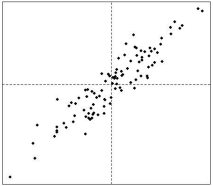
$$\rho_{Y_1 Y_2} = \frac{E\left[(Y_1 - E(Y_1))(Y_2 - E(Y_2))\right]}{\sigma_{Y_1}\sigma_{Y_2}}$$

Sample estimator:

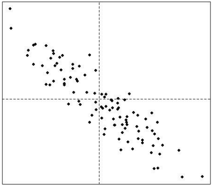
$$r_{Y_1 Y_2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)}{s_{Y_1} s_{Y_2}}$$



$r = 0$



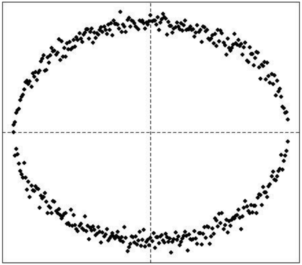
$r > 0$



$r < 0$

$\rho = 0$ does not imply lack of association

(unless Y_1 and Y_2 follow a bivariate normal distribution)



$r = 0.003$

Example

As an example, we will use the data set `trees` in *R*. This data set provides measurements of the girth, height and volume of timber in 31 felled black cherry trees. Note that girth is the diameter of the tree (in inches) measured at 4 ft. 6 in. above the ground. For this data set, `Volume` is the *response variable*, and `Height` and `Girth` are the *explanatory variables*.

```
> data(trees)
> pairs(trees)
```

The graph shows a strong association between Volume and Girth, and some degree of association between Height and Volume. The explanatory variables also seem to be associated.

The command in *R* for obtaining correlations is

```
cor(x, y = NULL, method = c("pearson",
    "kendall", "spearman"))
```

where `x` can be a numeric vector, matrix or data frame, `y` is 'NULL' (default) or a vector, matrix or data frame with compatible dimensions to 'x (the default is equivalent to `'y = x'`,but more efficient), and `method` is a character string indicating which correlation coefficient will be calculated.

```
> cor(trees)
           Girth    Height    Volume
Girth  1.0000000 0.5192801 0.9671194
Height 0.5192801 1.0000000 0.5982497
Volume 0.9671194 0.5982497 1.0000000
```

The estimated correlations are compatible with the observations from the graphics: Volume and Girth seem to be strongly associated, Volume and Height also have some degree of linear association, but as Height and Girth are also associated, there is the doubt if the association between Volume and Height is due to the association between the explanatory variables.

Linear Models

In the following discussion, we will consider models of the form

$$\text{Observation} = \text{signal} + \text{noise}$$

Considering again the black cherry trees example, suppose that we want to predict the volume of timber of a tree using the girth and the height of that tree. It seems natural to establish a relationship like the following:

$$\text{Volume}_i = \beta_0 + \beta_1 \text{Girth}_i + \beta_2 \text{Height}_i + \varepsilon_i$$

Linear Models

In the following discussion, we will consider models of the form

$$\text{Observation} = \text{signal} + \text{noise}$$

Considering again the black cherry trees example, suppose that we want to predict the volume of timber of a tree using the girth and the height of that tree. It seems natural to establish a relationship like the following:

$$\text{Volume}_i = \beta_0 + \beta_1 \text{Girth}_i + \beta_2 \text{Height}_i + \varepsilon_i$$

Consider another example

Genetically similar seeds are randomly assigned to be raised in either a nutritionally enriched environment (treatment group) or standard conditions (control group) using a completely randomized experimental design. After a predetermined time all plants are harvested, dried and weighed.

It seems reasonable to establish that for each group (control and treatment) there is a central value around which observations are distributed, and that value is different for both groups. An adequate model, then, could be

$$Y_{j,k} = \mu_j + \varepsilon_{jk}$$

where j is the environment and k is the plant.

The examples considered are classified as *linear models*, because they are linear in the parameters, and the noise is additive.

Basic hypothesis on the errors:

1. The error have mean 0: $E[\varepsilon_i] = 0$
2. The variance of the errors is constant, σ^2: $Var[\varepsilon_i] = \sigma^2$
3. Errors are independent.

In particular, we will work with $\varepsilon_i \sim N(0, \sigma^2)$ (equivalently, $\varepsilon \sim N_n(\mathbf{0}, \sigma^2 I)$).

A linear model of the form $y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i$ is called *regression model*.

The examples considered are classified as *linear models*, because they are linear in the parameters, and the noise is additive.

Basic hypothesis on the errors:

1. The error have mean 0: $E[\varepsilon_i] = 0$
2. The variance of the errors is constant, σ^2: $Var[\varepsilon_i] = \sigma^2$
3. Errors are independent.

In particular, we will work with $\varepsilon_i \sim N(0, \sigma^2)$ (equivalently, $\varepsilon \sim N_n(\mathbf{0}, \sigma^2 I)$).

A linear model of the form $y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i$ is called *regression model.*

The examples considered are classified as *linear models*, because they are linear in the parameters, and the noise is additive.

Basic hypothesis on the errors:

1. The error have mean 0: $E[\varepsilon_i] = 0$
2. The variance of the errors is constant, σ^2: $Var[\varepsilon_i] = \sigma^2$
3. Errors are independent.

In particular, we will work with $\varepsilon_i \sim N(0, \sigma^2)$ (equivalently, $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I)$).

A linear model of the form $y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i$ is called *regression model*.

The examples considered are classified as *linear models*, because they are linear in the parameters, and the noise is additive.

Basic hypothesis on the errors:

1. The error have mean 0: $E[\varepsilon_i] = 0$
2. The variance of the errors is constant, σ^2: $Var[\varepsilon_i] = \sigma^2$
3. Errors are independent.

In particular, we will work with $\varepsilon_i \sim N(0, \sigma^2)$ (equivalently, $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I)$).

A linear model of the form $y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i$ is called *regression model*.

Simple Regression Model

Consider a set of $i = 1$ to n observations with fixed X-values and random Y-values. The *simple regression model* is

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n \\ \varepsilon_i &\sim N(0, \sigma^2) \end{aligned}$$

This model can also be written as

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

The Y_i's are considered independent.

Simple Regression Model

Consider a set of $i = 1$ to n observations with fixed X-values and random Y-values. The *simple regression model* is

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n \\ \varepsilon_i &\sim N(0, \sigma^2) \end{aligned}$$

This model can also be written as

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

The Y_i's are considered independent.

Interpretation of the elements in the model:

- y_i is the value of Y for the ith observation when the predictor variable $X = x_i$.
- β_0 is the population intercept, the mean value of the probability distribution of Y when x_i equals zero.
- β_1 is the population slope and measures the change in Y per unit change in X.
- ε_i is a random or unexplained error associated with the ith observation.

Interpretation of the elements in the model:

- y_i is the value of Y for the ith observation when the predictor variable $X = x_i$.
- β_0 is the population intercept, the mean value of the probability distribution of Y when x_i equals zero.
- β_1 is the population slope and measures the change in Y per unit change in X.
- ε_i is a random or unexplained error associated with the ith observation.

Interpretation of the elements in the model:

y_i is the value of Y for the ith observation when the predictor variable $X = x_i$.

β_0 is the population intercept, the mean value of the probability distribution of Y when x_i equals zero.

β_1 is the population slope and measures the change in Y per unit change in X.

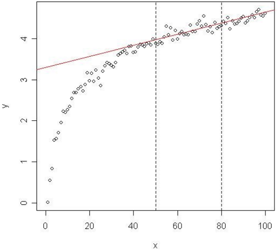
ε_i is a random or unexplained error associated with the ith observation.

Interpretation of the elements in the model:

- y_i is the value of Y for the ith observation when the predictor variable $X = x_i$.
- β_0 is the population intercept, the mean value of the probability distribution of Y when x_i equals zero.
- β_1 is the population slope and measures the change in Y per unit change in X.
- ε_i is a random or unexplained error associated with the ith observation.

A comment on the interpretation of β_0

Linear models are, in many cases, approximations to the real situation. When observed X-values are very far from zero, β_0 can't be interpreted as $E[Y|X = 0]$, but rather as a structural parameter in the model which allows a good *local* approximation of the relationship under study.



Estimation in the Simple Regression Model

The problem of estimating β_0 and β_1 is equivalent to the problem of adjusting the "best fitting line" to the data under study.

Call $\hat{y}_i$ the value of y_i predicted by the fitted regression line for each x_i, that is,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Define the *residuals* $e_i = y_i - \hat{y}_i$. Then, a criterion for selecting the best fitting line is taking $\hat{\beta}_0$ and $\hat{\beta}_1$ such that minimize

$$SSE = \sum_{i=1}^{n} e_i^2$$

This procedure is called *Least Squares Method*.

Estimation in the Simple Regression Model

The problem of estimating β_0 and β_1 is equivalent to the problem of adjusting the "best fitting line" to the data under study.
Call $\hat{y}_i$ the value of y_i predicted by the fitted regression line for each x_i, that is,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Define the *residuals* $e_i = y_i - \hat{y}_i$. Then, a criterion for selecting the best fitting line is taking $\hat{\beta}_0$ and $\hat{\beta}_1$ such that minimize

$$SSE = \sum_{i=1}^{n} e_i^2$$

This procedure is called *Least Squares Method*.

Estimation in the Simple Regression Model

The problem of estimating β_0 and β_1 is equivalent to the problem of adjusting the "best fitting line" to the data under study.
Call $\hat{y}_i$ the value of y_i predicted by the fitted regression line for each x_i, that is,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Define the *residuals* $e_i = y_i - \hat{y}_i$. Then, a criterion for selecting the best fitting line is taking $\hat{\beta}_0$ and $\hat{\beta}_1$ such that minimize

$$SSE = \sum_{i=1}^{n} e_i^2$$

This procedure is called *Least Squares Method*.

Estimation in the Simple Regression Model

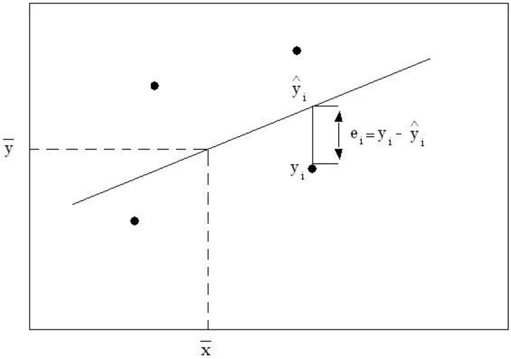
The problem of estimating β_0 and β_1 is equivalent to the problem of adjusting the "best fitting line" to the data under study.
Call $\hat{y}_i$ the value of y_i predicted by the fitted regression line for each x_i, that is,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Define the *residuals* $e_i = y_i - \hat{y}_i$. Then, a criterion for selecting the best fitting line is taking $\hat{\beta}_0$ and $\hat{\beta}_1$ such that minimize

$$SSE = \sum_{i=1}^{n} e_i^2$$

This procedure is called *Least Squares Method*.



$\bar{y}$
$\hat{y}_i$
$e_i = y_i - \hat{y}_i$
y_i
$\bar{x}$

Parameter	LS estimate	Standard Error
β_1	$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$	$\sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$
β_0	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$	$\sqrt{MSE \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$
ε_i	$e_i = y_i - \bar{y}_i$	$\sqrt{MSE}$

$MSE = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$ is the estimator of σ^2.

Note that $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combinations of the y_i's, and so the have normal distribution. It can also be shown that they are unbiased estimators of β_0 and β_1.

In *R*, a linear regression model can be adjusted using the command `lm(y~x)`.
Suppose we want to fit a simple linear regression model which explains the volume of timber of black cherry trees using girth as explanatory variable

```
> trees.mod1 = lm(Volume~Girth)
> names(trees.mod1)
 [1] "coefficients"  "residuals"     "effects"
 [4] "rank"          "fitted.values" "assign"
 [7] "qr"            "df.residual"   "xlevels"
[10] "call"          "terms"         "model"
```

The output of `lm` is a list.

```
> trees.mod1$coeff
(Intercept)       Girth
 -36.943459    5.065856
```

Tests of Hypothesis on the Parameters of the Model

It can be shown that, under the assumption of normality of the errors,

$$t_i = \frac{\hat{\beta}_i - \beta i}{\text{Standard Error}(\hat{\beta}_i)} \sim t_{n-2}$$

So, if we want to test the hypothesis

$$H_0 : \quad \beta_i = 0 \quad \text{vs.} \;\; \beta_i \neq 0$$

the test can be based on the statistic

$$t = \frac{\hat{\beta}_i}{\text{Standard Error}(\hat{\beta}_i)}$$

If H_0 is true, $t \sim t_{n-2}$. Then, we will reject H_0 if the absolute value of the observed t statistic, t_{obs}, is greater than $t_{n-2,\frac{\alpha}{2}}$. So, our rejection region will be

$$\left|t_{obs}\right| > t_{n-2,\frac{\alpha}{2}}$$

For getting those tests in *R*, it is enough to use

```
Call:
lm(formula = Volume ~ Girth)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0654 -3.1067  0.1520  3.4948  9.5868

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
Girth         5.0659     0.2474   20.48  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 4.252 on 29 degrees of freedom
Multiple R-Squared: 0.9353,     Adjusted R-squared: 0.9331
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

Note that $\hat{\beta}_0$ is negative, and so it can't be interpreted as the volume of timber when the girth approaches zero!

When we test the hypothesis $\beta_1 = 0$, the p-value is $< 2 \times 10^{-16}$. This means that in this example the data discredits the null hypothesis so we can reject it. Our conclusion, then, is that the volume of timber can be predicted using the girth as explanatory variable.

Analysis of Variance

Define the **total sum of squares**, SST, as

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

SST can be seen as the variation in the data which can't be explained just using the simple model $y_i = \beta_0 + \varepsilon$ (for this model, $\hat{\beta}_0 = \bar{y}$)

As only one parameter (β_0) has been estimated, we say that SST has $n - 1$ degrees of freedom

SST can be partitioned in two additive components

$$SST = SSR + SSE$$

where

- $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ is the *regression sum of squares*, which can be interpreted as the amount of variation explained by the explanatory variable.
  $df_{Regression} = 1$
- $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the sum of the squared errors or *residual sum of squares*, which can be interpreted as the variation in Y not explained by the regression.
  $df_{Residual} = n - 2$

These sums of squares are usually presented in a *Analysis of Variance (ANOVA)* table

Source	df	SS	MS
Regression	1	$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$MSR = \frac{SSR}{1}$
Residual	$n-2$	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$MSE = \frac{SSE}{n-2}$
TOTAL	$n-1$	$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$	

It can be shown that

$$\begin{aligned} E(MSR) &= \sigma_\varepsilon^2 + \beta_1 \sum_{i=1}^{n} (x_i - \bar{x}_i) \\ E(MSE) &= \sigma_\varepsilon^2 \end{aligned}$$

- **Observation 1** *MSE* is an unbiased estimator of σ_ε^2
- **Observation 2** When $\beta_1 = 0$, *MSR* is also an estimator of σ_ε^2.

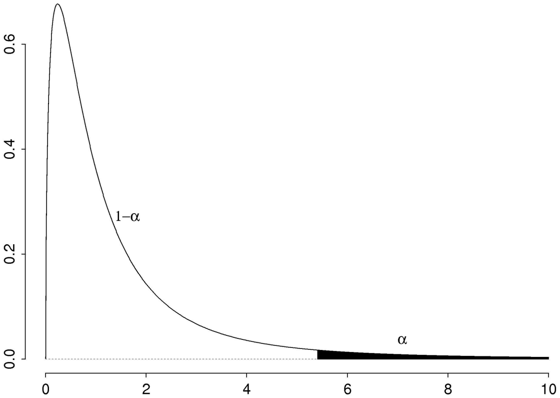
Observations 1 and 2 imply that, when $\beta_1 = 0$, the quotient $\frac{MSR}{MSE}$ must be near 1. It can also be shown that, when $\beta_1 = 0$

$$\frac{MSR}{MSE} \sim F_{1,n-2}$$

This can be used for testing $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$. We will reject H_0 with level α when

$$F = \frac{MSR}{MSE} > F^{\alpha}_{1,n-2}$$

Rejection region for the F test



In our example, we obtained

```
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

As with the *t*-test, the null hypothesis $\beta_1 = 0$ can be rejected.
The ANOVA table for the model can be extracted using the command `anova`

```
> anova(trees.mod1)
Analysis of Variance Table

Response: Volume
          Df Sum Sq Mean Sq F value    Pr(>F)
Girth      1 7581.8  7581.8  419.36 < 2.2e-16 ***
Residuals 29  524.3    18.1
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Multiple Determination Coefficient (R^2)

A common measure used for study the adequacy of the model is the *Multiple Determination Coefficient* R^2, defined as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

R^2 could be informally interpreted as the percentage of the total variation which is explained by the model. Values of R^2 close to 1 indicate that the model is providing a good fit for the data.

Note that R^2 does not take into account neither the size of the data nor the complexity of the model. The *adjusted multiple determination coefficient*, R^2_{adj}, is defined as

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-2}}{\frac{SST}{n-1}}$$

The importance of this complexity adjustment will be evident when we study more complex models.

In our example, $R^2 = 0.9353$ and $R^2_{adj} = 0.9331$. Both numbers indicate that the model with `Girth` as explanatory variable provides a good fit for `Volume`.

Regression Diagnostics

The fitting and analysis of a linear model is based on four basic assumptions

- The relationship between variables is linear.
- Errors are normally distributed.
- Error variances are homogeneous (*homocedasticity*)
- Errors are independent.

It is important to check if these assumptions hold.

Residuals

We defined residuals as

$$e_i = y_i - \hat{y}_i$$

Patterns in residuals are an important way of checking regression assumptions.

The variance of sample residuals may not be constant for different x_i's, in contrast with model error terms ε_i. *Standarized residuals* (called in some texts *Studentized*) should have constant variance if the model is correct.

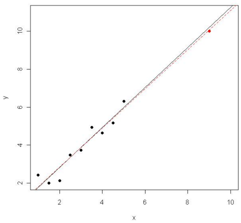
$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_i)}}$$

Leverage

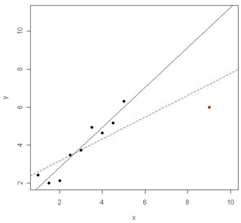
In last equation, the *leverage* h_i measures how influential is observation i according to its x value.

In simple linear regression, $h_i = \frac{1}{n} + \frac{x_i - \bar{x}}{\sum_{k=1}^{n}(x_k - \bar{x})^2}$. The higher the distance between x_i and $\bar{x}$, the higher the leverage.
When an observation with high leverage does not follow the same linear pattern than the other observations, it can change the fitted line. We say that it is an *influential* observation.

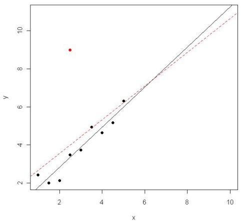
(1) High leverage, low residual



(2) High leverage, high residual



(3) Low leverage, high residual



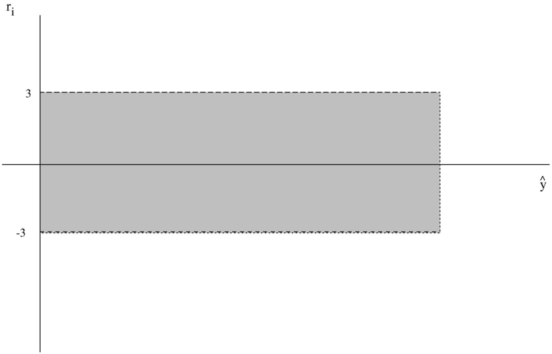
Red and black lines are fitted with and without the red point, respectively. In (2) and (3), red points are influential, even though the point in (3) has low leverage.

Coook's distance measures the effect of the i-th observation on the estimation of coefficients. A large Cook's distance identifies an influential observation.

Residual plots

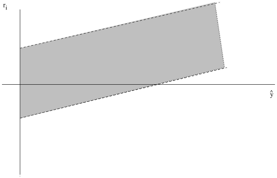
- **Normal probability plot of standarized residuals**
  It should be similar to a straight line. Is very useful for detecting outliers (observations with unusually large residuals)
- **Residuals vs explanatory variables plot**
- **Residuals vs fitted values plot**
  If the model fits well the data, we would expect that residuals contain the "noise" part of the model, and that there are not patterns in them.

Form of a satisfactory residuals vs fitted values plot

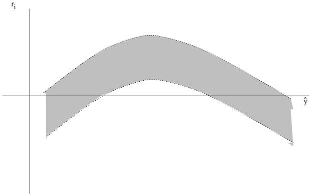


Typical forms of non-satisfactory residual plots

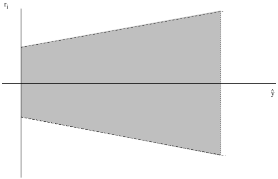
(1) Missing linear term



(2) Missing higher order term



(3) Heterocedasticity

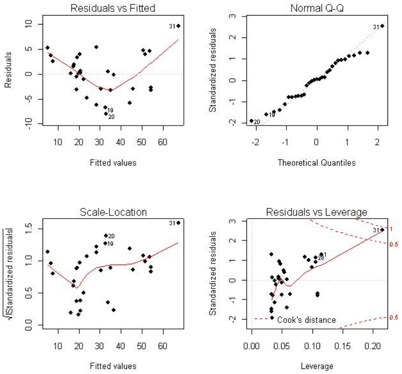


Applying `plot` to a linear model object gives four plots as output

- Plot of residuals vs. fitted values.
- Normal plot of standarized residuals.
- Scale-location plot .

  This plot (also called “Spread-Location”) takes the square root of the absolute residuals in order to diminish skewness. If the model is adequate, we will expect all points near to zero. Allows identification of specially large residuals.
- Plot of residual vs leverage, with level curves for Cook’s distances.

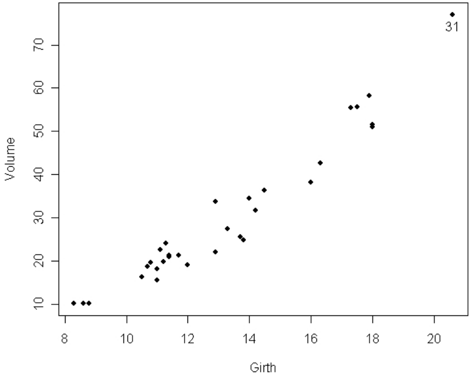
```
> par(mfrow=c(2,2))
> plot(trees.mod1,pch=16)
```



- The residuals vs fitted values plot has a curved form (positive residuals in the extremes, negative residuals at the center). This suggests that we should adjust a different model (for example, $\text{Volume}_i = \beta_0 + \beta_1 \text{Girth}_i^2 + \varepsilon_i$)
- Observation 31 is an influential observation, and it seems to be an outlier with respect to the model (an observation with an unusually large residual)

For identifying observation 31, we can use

```
> par(mfrow=c(1,1))
> plot(Girth,Volume,pch=16)
> identify(Girth,Volume,labels=1:31)
```



Multiple Linear Regression

In most cases, we will need more than one explanatory variable in order to explain the variability in the data. A *multiple regression model* is a linear model with the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

The assumptions on the errors are the same as before.
The estimation of the parameters is done through least squares. The estimators such obtained are unbiased and follow a normal distribution.
The variance of the error is estimated as

$$\hat{\sigma}^2 = MSE = \frac{\sum e_i^2}{n-(k+1)}$$

where $e_i = y_i - \hat{y}_i$ is the residual of the $i-th$ observation.

Hypothesis Testing for the Model Parameters

We want to test: $H_0 : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$.

This is equivalent to compare the models

$$y_l = \beta_0 + \beta_1 x_{1l} + \ldots + \beta_{i-1} x_{i-1,l} + \beta_{i+1} x_{i+1,l} + \ldots + \beta_k x_{kl} + \varepsilon_l$$

vs. the model

$$y_l = \beta_0 + \beta_1 x_{1l} + \ldots + \beta_{i-1} x_{i-1,l} + \beta_i x_{i,l} + \beta_{i+1} x_{i+1,l} + \ldots + \beta_k x_{kl} + \varepsilon_l$$

As in the simple regression case, it can be shown that

$$\frac{\hat{\beta}_i - \beta_i}{\text{Standard error } \hat{\beta}_i} \sim t_{n-p},$$

where $p = k + 1$ is the total number of parameters in the model.
If H_0 is true

$$t = \frac{\hat{\beta}_i}{\text{Standard error } \hat{\beta}_i} \sim t_{n-p},$$

We reject H_0 when

$$|t| > t_{n-p,\alpha/2}$$

The t-test can lead to wrong results if it is not used carefully, as the $\hat{\beta}_i$ are not independent

In general, it is not advisable to remove more than one variable at a time applying this procedure, as the t-test only allows to compare models differing in one variable.

In the black cherry trees example, we have

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.92041   10.07911  -0.984 0.333729
Girth       -2.88508    1.30985  -2.203 0.036343 *
I(Girth^2)   0.26862    0.04590   5.852 3.13e-06 ***
Height       0.37639    0.08823   4.266 0.000218 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The null hypothesis $H_0 : \beta_i = 0$ can be rejected for all variables in the model. Then, it seems that this model cannot be simplified.

Comparing Linear Models

We wish to compare models

$$\begin{aligned}
(1)\ \ y_i &= \beta_0 + \beta_1 x_{1i} + \ldots + \beta_l x_{li} + \varepsilon_i \\
(2)\ \ y_i &= \beta_0 + \beta_1 x_{1i} + \ldots + \beta_l x_{li} + \beta_{l+1} x_{(l+1)i} + \ldots + \beta_k x_{ki} + \varepsilon_i
\end{aligned}$$

All explanatory variables of model (1) are contained in model (2): model (1) is *nested* in model (2).

Comparing these models is equivalent to testing the hypotheses

$$\begin{aligned}
H_0: &\quad \beta_{l+1} = \beta_{l+2} = \ldots = \beta_k = 0 \ \text{ vs.} \\
H_1: &\quad \beta_j \neq 0, \text{some } j = l+1, \ldots, k
\end{aligned}$$

It can be shown that

$$\frac{1}{\sigma^2}(SSE_1 - SSE_2) \sim \chi^2_{k-l}$$

$$\frac{1}{\sigma^2}SSE_2^2 \sim \chi^2_{n-(k+1)}$$

where SSE_i is the sum of the squares of the residuals for model i. It can also be proven that this random variables are independent.

Using this result,

$$F = \frac{\frac{SSE_1 - SSE_2}{k-l}}{\frac{SSE_2}{n-(k+1)}} \sim F_{k-l,n-(k+1)}$$

We reject H_0 if $F > F^{\alpha}_{k-l,n-(k+1)}$

Particular case: Test of significance of the model

$$
\begin{aligned}
H_0 : &\quad \beta_1 = \beta_2 = \ldots = \beta_k = 0 \ \text{ vs.} \\
H_1 : &\quad \beta_j \neq 0, \text{some } j = 1, \ldots, k
\end{aligned}
$$

The F statistic for this test has the form

$$
\begin{aligned}
F &= \frac{\frac{(\sum_{i=1}^n y_i^2 - n\bar{y}^2) - (\sum_{i=1}^n y_i^2 - \sum_{i=1}^n \hat{y}_i^2)}{k}}{\frac{\sum_{i=1}^n y_i^2 - \sum_{i=1}^n \hat{y}_i^2}{n-(k+1)}} \\
&= \frac{\frac{\sum_{i=1}^n \hat{y}_i^2 - n\bar{y}^2}{k}}{\frac{\sum_{i=1}^n y_i^2 - \sum_{i=1}^n \hat{y}_i^2}{n-(k+1)}}
\end{aligned}
$$

Denoting $SSR = \sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}^2$ (regression sum of squares), we can write

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-(k+1)}} = \frac{MSR}{MSE}$$

All the information for the test of significance of the model is contained in the *Analysis of Variance Table* or *ANOVA Table*

Source	df	SS	MS	F
Model	k	$SSR = \sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}^2$	$MSR = \frac{SSR}{k}$	$\frac{MSR}{MSE}$
Error	$n-(k+1)$	$SSE = \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} \hat{y}_i^2$	$MSE = \frac{SSE}{n-(k+1)}$	
TOTAL	$n-1$	$SST = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$		

The `summary` of a linear model object includes the F value corresponding to testing the significance of the model and the corresponding p-value

```
Residual standard error: 2.625 on 27 degrees of freedom
Multiple R-Squared: 0.9771,     Adjusted R-squared: 0.9745
F-statistic: 383.2 on 3 and 27 DF,  p-value: < 2.2e-16
```

In this particular example, the value of F allows to reject $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. This means that the explanatory variables are in fact explaining the variability of the data.

In *R*, we can use the command `anova` for obtaining the ANOVA table for comparing models. In the case of the black cherry trees data

```
> trees.mod0 = lm(Volume ~ 1, data=trees)
> summary(trees.mod0)

Call: lm(formula = Volume ~ 1, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-19.971 -10.771  -5.971   7.129  46.829

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   30.171      2.952   10.22 2.75e-11 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 16.44 on 30 degrees of freedom
```

```
> anova(trees.mod0,trees.mod1)
Analysis of Variance Table

Model 1: Volume ˜ 1
Model 2: Volume ˜ Girth + I(Girthˆ2) + Height
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     30 8106.1
2     27  186.0  3    7920.1 383.20 < 2.2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The F-statistic and the p-value are the ones we already analyzed.

The command `anova` can be applied to just one model. In that case, the explanatory variables are tested in the order they were added to the model.

```
> anova(trees.mod1)
Analysis of Variance Table

Response: Volume
           Df Sum Sq Mean Sq  F value    Pr(>F)
Girth       1 7581.8  7581.8 1100.511 < 2.2e-16 ***
I(Girth^2)  1  212.9   212.9   30.906 6.807e-06 ***
Height      1  125.4   125.4   18.198 0.0002183 ***
Residuals  27  186.0     6.9
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
> trees.mod1r=lm(Volume~Height+Girth+I(Girth^2),data=trees)
> anova(trees.mod1r)
Analysis of Variance Table

Response: Volume
           Df Sum Sq Mean Sq F value    Pr(>F)
Height      1 2901.2  2901.2 421.113 < 2.2e-16 ***
Girth       1 4783.0  4783.0 694.259 < 2.2e-16 ***
I(Girth^2)  1  235.9   235.9  34.243  3.13e-06 ***
Residuals  27  186.0     6.9
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Multiple Determination Coefficient R^2

We can define the *Multiple Determination Coefficient* R^2, in a way similar to that we used for the simple linear regression

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Note that R^2 doesn't penalize model complexity (every time a new explanatory variable is introduced in the model, R^2 increases).

To solve that problem, we can again define the *Adjusted Multiple Determination Coefficient*, R^2_{adj}, as

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-(k+1)}}{\frac{SST}{n-1}} = 1 - \frac{MSE}{MST}$$

Models with the same SSE can generate different values of R^2_{adj}, depending on the commplexity of the model.

For the black cherry trees data

```
Residual standard error: 2.625 on 27 degrees of freedom
Multiple R-Squared: 0.9771,     Adjusted R-squared: 0.9745
F-statistic: 383.2 on 3 and 27 DF,  p-value: < 2.2e-16
```

According to its R^2 value, the model is explaining a 97% of the variation of the data. Note that R^2_{adj} is almost equal to R^2, and we have no reasons to suspect that the model contains superfluous variables.

As for simple linear regression, the residuals of the model should be analyzed

```
> par(mfrow=c(2,2))
> plot(trees.mod1)
```

The residuals do not give evidence of problems in the fitting.

Prediction

If $\mathbf{x}^* = (1, x_1^*, x_2^*, ..., x_k^*)$ is a vector of specific values of the explanatory variables, the fitted expected value of the response in $\mathbf{x}^*$ is

$$\hat{y}^* = \mathbf{x}_h' \hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_k x_k^*$$

In the case of the simple regression model, it can be shown that $\hat{y}^*$ is normally distributed with mean $y^* = \beta_0 + \beta_1 x^*$ and variance $\sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$. The standard error of this fitted value is, then

$$se.fit(\hat{y}^*|x) = \sqrt{MSE \left[\frac{1}{n} + \frac{(x^* - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}$$

A $100 \times (1 - \alpha)\%$ confidence interval for $m = E(y|x^*)$ can be calculated as

$$\hat{y}^* - t_{\alpha/2}^{n-2} se.fit(\hat{y}^*|x) < m < \hat{y}_h + t_{\alpha/2}^{n-2} se.fit(\hat{y}^*|x)$$

Note that when x^* is far from $\bar{x}$, the variance of the estimates increase. If what we want is to *predict* a future value of y when $x = x_f$, a wider interval (taking into account also the variance of the process) is obtained, as

$$se.pred(\hat{y}_f^*|x_f) = \sqrt{MSE \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]}$$

The command for obtaining predicted values in *R* is

```
predict[.lm](object, newdata, se.fit = FALSE,
             interval = c("none", "confidence", "prediction"),
             level = 0.95, type = c("response", "terms"),...)
```

where

- `object` is the model from which we want to predict
- `newdata` An optional data frame in which to look for variables with which to predict. If omitted, the fitted values are used.
- `se.fit` A switch indicating if standard errors are required..
- `interval` indicates if a confidence interval or a prediction interval should be calculates (by default, no interval is calculated)
- `level` Tolerance/confidence level for the intervals

For the black cherry trees data, lets predict the volume of timber obtained from a ”mean” tree

```
> attach(trees)
> mean(Girth)
[1] 13.24839
> mean(Height)
[1] 76
> predict(trees.mod1,data.frame(Girth=13.25,Height=76),
+ interval="confidence")
          fit      lwr      upr
[1,] 27.61776 26.29796 28.93756
> predict(trees.mod1,data.frame(Girth=13.25,Height=76),
+ interval="prediction")
          fit      lwr      upr
[1,] 27.61776 22.07285 33.16267
```

Comparing non-nested models

The F-test only allows comparing nested models. There are no hypothesis tests for comparing non-nested models for the same data set, but we can use several criteria, most based on adjusting SSE by model complexity.

Some criteria of this class are

- R^2 and its adjusted by complexity version R^2_{adj}
- The *Akaike Information Criterion (AIC)*

  The most common definition of AIC is

  $$AIC = -2(\text{maximum likelihood}) + 2(\text{number of parameters})$$

  For a regression model with n observations, p parameters and normally distributed errors with unknown common variance,

  $$AIC = n\log(SSE/n) + 2p$$

  The "best" model will be that with the lower AIC

- The *Bayesian Information Criterion (BIC) or Schwartz Criterion* is similar to AIC, but changes the weight for model complexity:

$$BIC = -2(\text{maximum likelihood}) + \log(n)(\text{number of parameters})$$

For a regression model with n observations, p parameters and normally distributed errors with unknown common variance,

$$BIC = n\log(SSE/n) + p\log(n)$$

- Mallows' C_p is defined as $C_p = \frac{SSE}{\hat{\sigma}^2} + 2p - n$, where $\hat{\sigma}^2$ is from the model with all predictors and SSE is for the model with p parameters. When all p parameters are used in the model, $C_p = p$. A model with a bad fit will produce a C_p much bigger than p. Desirable models have small p and C_p lower or equal than p.

For practicing all the concepts we have studied on regression analysis, lets analyze the data frame `swiss` in *R*. These data contain measures of fertility and socio-economic indicators for each of the 47 French speaking provinces in Switzerland around 1888.

Socio-economic indicators

- `Agriculture:` Percentage of males involved in agriculture as occupation.
- `Examination:` Percentage of "draftees" receiving highest mark on army examination.
- `Education:` Percentage of education beyond primary school for "draftees".
- `Catholic:` Percentage of catholic (as opposed to "protestant")
- `Infant.Mortality:` Percentage of live births who live less than 1 year

Response variable is an standarized fertility index, `Fertility`

```
> data(swiss)
> attach(swiss)
> pairs(swiss)
```

Fertility seems to have positive linear relationship with Agriculture and with Infant.Mortality, and negative linear relationship with Examination and Education. Nevertheless, it has to be taken into account that the explanatory variables are correlated.

In what follows we will use the following new commands

- add1: Given a model and a set of extra variables, this command determines the change in the residual sum of squares and in the AIC produced by the inclusion of each extra variable into the original model.

```
> swiss.mod0=lm(Fertility ~1,data=swiss)
> add1(swiss.mod0,.~Agriculture+Examination+Education
+ +Catholic+Infant.Mortality)
Single term additions
Model:
Fertility ~ 1
                 Df Sum of Sq    RSS    AIC
<none>                        7178.0  238.3
Agriculture       1     894.8 6283.1  234.1
Examination       1    2994.4 4183.6  215.0
Education         1    3162.7 4015.2  213.0
Catholic          1    1543.3 5634.7  229.0
Infant.Mortality  1    1245.5 5932.4  231.4
```

- drop1: Is equivalent to add1, but now eliminating variables.

```
> swiss.mod12345 =lm(Fertility ~ Agriculture+Examination+Education
+ +Catholic+Infant.Mortality,data=swiss)
> drop1(swiss.mod12345)
Single term deletions
Model:
Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality
                 Df Sum of Sq    RSS   AIC
<none>                        2105.0 190.7
Agriculture       1     307.7 2412.8 195.1
Examination       1      53.0 2158.1 189.9
Education         1    1162.6 3267.6 209.4
Catholic          1     447.7 2552.8 197.8
Infant.Mortality  1     408.8 2513.8 197.0
```

It is practically equivalent to eliminate variables using the t-test

- `step`: performs an automatic process of selection based on the AIC, either adding variables (`direction = ‘‘forward’’`), droping them (`direction= ‘‘backward’’`) or both (`direction = ‘‘both’’`) Default option is doing both.

```
> step(swiss.mod0,.~Agriculture+Examination+Education
+ +Catholic+Infant.Mortality)
```

- The package `leaps` in R contains the function `regsubsets()`, which is very useful for computing, for example, the adjusted R^2 and Mallows' C_p

```
> library(leaps)
> swiss.rss=regsubsets(Fertility~.,data=swiss)
> summary(swiss.rss)$adjr2
> summary(swiss.rss)$cp
```